NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

Statistical Techniques Research Group
Section of Mathematical Statistics
Department of Mathematics
Princeton University, Princeton, N.J.

Technical Report No. 42
July 1961

THE VARIANCE OF MEANS OF SYMMETRICALLY TRIMMED SAMPLES FROM NORMAL POPULATIONS, AND ITS ESTIMATION FROM SUCH TRIMMED SAMPLES. (TRIMMING/WINSORIZATION I)

by

Donald H. McLaughlin and John W. Tukey

Department of Army Project No. 5B 99-01-004
Ordnance R and D Project No. PB2-0001
OOR Contract No. 1715
Contract No. 36-034-ORD-2297

1. Summary.

The use of the mean of a symmetrically trimmed sample (the trimmed mean) as an indicator of location and the use of the total sum of squares of deviations of the same trimmed sample (the TSSD) as an indicator of the variability of the trimmed mean are explored. The increase in variance (of the trimmed mean as compared with the untrimmed mean) when trimming samples from *an exactly normal distribution* is found to be less than 3%, 6%, 9%, and 14%, respectively, when a total of 1/10, 2/10, 3/10, or 4/10 of the sample is trimmed away. (Trimming will *decrease* the variance when the samples come from a long-tailed distribution.) The loss of normal-theory efficiency is given for all symmetric trimmings of samples of size ≤ 20. The appropriate divisor, by which the trimmed sum of squares of deviations is to be divided to obtain an estimate of the variance of the trimmed mean,is tabled for the same range.

The effect on this divisor of sampling from rectangular rather than normal populations is found to be small, but noticeable. The empirical behavior of the reciprocal of the divisor is found to be simple, and a theoretical explanation for this is provided.

Further studies in this area are in progress.

2. The problem.

While the sample mean and sample variance are sufficient statistics when the sample is specified to come from a precisely normally distributed population, so that no statistic can then be a better estimate of location than the sample mean, and no statistic can be a better basis for estimating the variance of the sample mean than the sample variance (or sample sum of squares of deviations), these optimum properties fail miserably for samples from non-normal distributions (even when these non-normal distributions are symmetrical). Thus it is of interest to consider other indicators of location, and other bases for estimating the variance of these indicators. We must decide how to work numerically with each particular indicator and with each particular basis for estimating its variance. The first requirement that our procedures must satisfy is that they be appropriate when the parent distribution is precisely normal. (Though we may rarely find samples from normal distributions in practice, none of us want to give up the possibility that a few of the parent populations we face may be almost exactly normal, and that others may be nearly normal.)

In large samples the trimmed mean and the trimmed standard deviation (the mean and sample standard deviation of those observed values remaining out of a sample of n when the g highest and g lowest values have been deleted) have been shown to have quite satisfactory properties [5]. This report opens an investigation into the properties of both these and related statistics in small and moderate samples.

3. Results.

The main quantities studied are:

(1) the variance of the trimmed mean.

(2) the normal theory efficiency for location of the trimmed mean (as referred to the untrimmed mean; i. e., the ratio of the variance of the untrimmed mean to the variance of the trimmed mean).

(3) the average value of the trimmed sum of squares of deviations.

(4) the ratio of (3) to (1), which is the factor by which an observed trimmed sum of squares of deviations should be divided in order to obtain an unbiased estimate of the variance of the corresponding trimmed mean. These values are given for $n \leq 20$.

Such numerical results provide (i) a method for calculating an unbiased estimate of the variance of a trimmed mean, and (ii) an indication of the price that must be paid for trimming when the parent population is indeed normal. They do not provide solutions for the following problems:

(1) If the extent of trimming is guided by the apparent quality of the estimates provided by differently trimmed means, how much will be the downward bias of the estimated variance of that trimmed mean which appears to have the least variance? (This bias is due to selection and arises when the variances of the various trimmed means are estimated as indicated below.)

(2) How much does the distribution of the ratio of trimmed mean to the square root of its estimated variance (based upon the trimmed sum of squares of deviations) differ from a suitable multiple of a suitable instance of Student's t? What multiple and what degrees of freedom are suitable?

(3) How variable is the trimmed sum of squares of deviations as a basis for estimating scale?

(4) How do trimmed means and trimmed sums of squares of deviations behave for parent distributions that are non-normal but symmetric?

It is hoped to provide at least partial answers to these questions in later reports of this series.

The most directly relevant and useful results are collected in Table 1, Figure 1, and Table 2. Table 1 shows the loss in normal theory efficiency when 1, 2, 3, ... observed values are deleted from each end of a sample. Figure 1 shows similar information in terms of the modified fraction of the observations deleted from each end. Table 2 contains the divisors which convert trimmed sums of squares of deviations into unbiased estimates of variances of trimmed means.

Table 1

Loss in normal theory efficiency for location = $\frac{\text{var } \overset{\cup}{y} - \text{var } \bar{y}}{\text{var } \bar{y}}$

(Trimmed mean referred to untrimmed mean)

Size of original sample	Number of observed values deleted at each end = $\frac{n - h}{2}$								
	1	2	3	4	5	6	7	8	9
3	34.6%								
4	19.3%								
5	13.5%	43.4%							
6	10.4%	28.9%							
7	8.5%	22.2%	47.3%						
8	7.1%	18.2%	34.5%						
9	6.1%	15.4%	28.1%	49.5%					
10	5.3%	13.3%	23.8%	38.3%					
11	4.7%	11.7%	20.6%	32.2%	50.9%				
12	4.3%	10.5%	18.3%	28.0%	41.0%				
13	3.9%	9.4%	16.3%	24.8%	35.3%	51.8%			
14	3.5%	8.6%	14.8%	22.2%	31.2%	43.0%			
15	3.2%	7.9%	13.5%	20.1%	28.1%	37.8%	52.5%		
16	3.0%	7.3%	12.4%	18.4%	25.5%	33.9%	44.6%		
17	2.8%	6.7%	11.5%	17.0%	23.3%	30.7%	39.7%	53.1%	
18	2.6%	6.3%	10.7%	15.7%	21.5%	28.2%	36.0%	45.9%	
19	2.4%	5.9%	10.0%	14.6%	19.9%	26.0%	33.0%	41.3%	53.5%
20	2.3%	5.5%	9.3%	13.7%	18.6%	24.1%	30.4%	37.8%	46.9%

Chart I: Loss in Normal Theory Efficiency for Location When Trimmed Mean Replaces Untrimmed Mean.

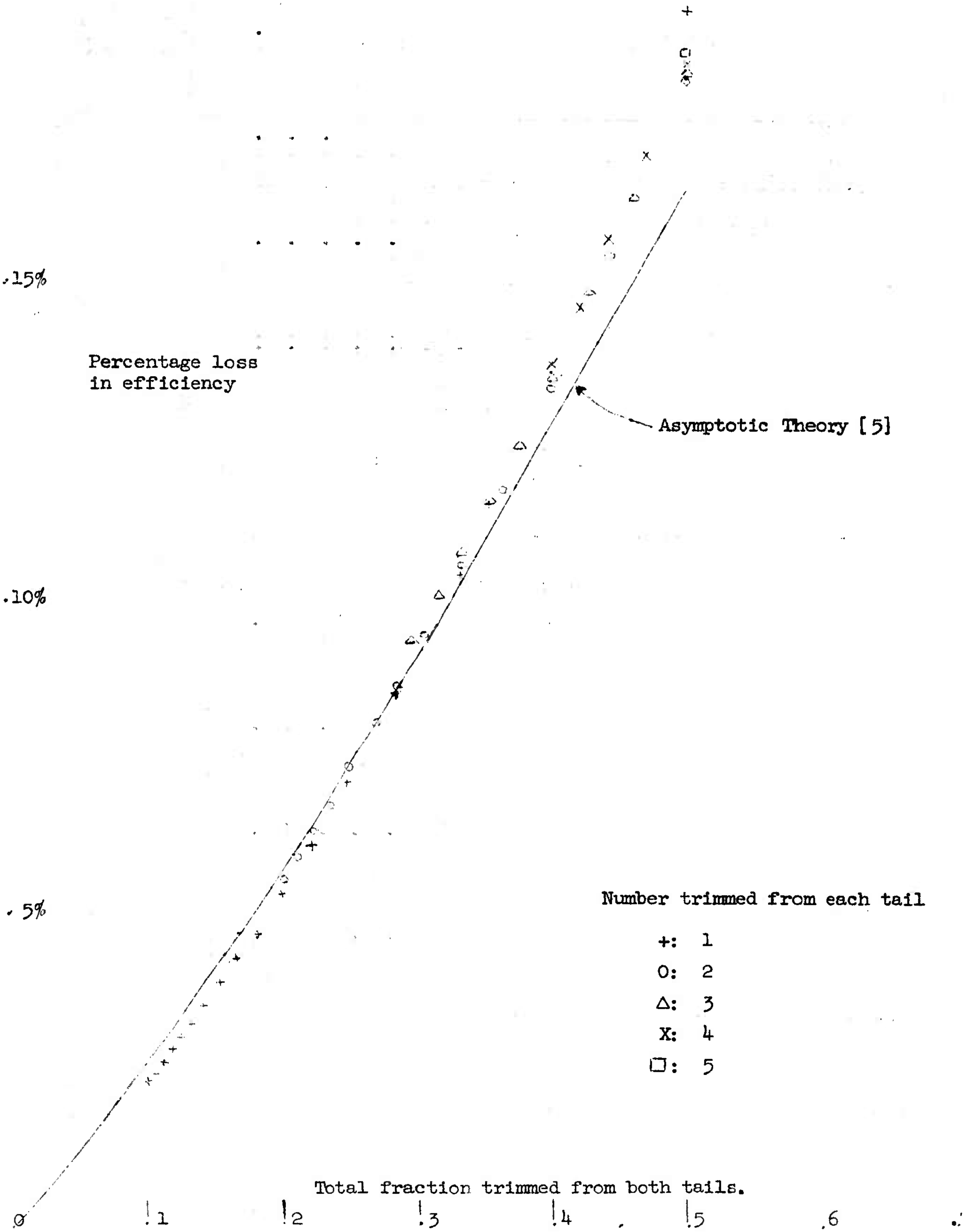

Table 2

Ratio of average sum of squares of deviations to variance of mean for trimmed samples from normal populations = factor by which a trimmed sum of squares of deviation is to be divided to obtain an unbiased (on normal theory, and for constant amount of trimming at each sample size) estimate of the variance of the corresponding trimmed means

size of sample after trimming	g = number of observations trimmed from each end									
h	0	1	2	3	4	5	6	7	8	9
2	2	1.0092	.67142	.50263	.40158	.33436	.28641	.25050	.22259	.20028
3	6	3.1657	2.1474	1.6243	1.3061	1.0921	.93833	.82253	.73217	
4	12	6.6313	4.5955	3.5181	2.3955	2.0659	2.0659	1.8160	1.6200	
5	20	11.519	8.1453	6.3090	5.1502	4.3514	3.7671	3.3212		
6	30	17.910	12.898	10.100	8.3053	7.0533	6.1296	5.4198		
7	42	25.866	18.935	14.979	12.401	10.583	9.2313			
8	56	35.436	26.323	21.016	17.511	15.014	13.142			
9	72	46.658	35.117	28.276	23.699	20.409				
10	90	59.563	45.364	36.811	31.024	26.827				
11	110	74.177	57.103	46.669	39.535					
12	132	90.523	70.370	57.891	49.276					
13	156	108.62	85.194	70.513						
14	182	128.48	101.60	84.569						
15	210	150.13	119.61							
16	240	173.57	139.26							
17	272	198.82								
18	306	225.88								
19	342									
20	380									

4. Example.

If we are dealing with samples of 11 and choose to routinely trim 2 observations off each end of each sample, the loss of normal efficiency can be seen from Table 1 to be 11.7%. If the population is exactly normal, the trimmed mean will have a standard deviation some 6% greater than the untrimmed mean. (And if the population has rather long tails, the trimmed mean will have a much smaller standard deviation than the untrimmed mean.)

If we have the following 11 observations: -5, 10, 15, 11, 12, 17, -1, 8, 13, 10, 18 and proceed by trimming two from each end, we have to find the mean and sum of squares of deviations of the remaining 7 observations. Hence

10	100
15	225
11	121
12	144
8	64
13	169
10	100
79	923

$$\overset{\sqcup}{y} = \frac{79}{7} = 11.28 = \text{trimmed mean}$$

$$T = 923 - \frac{79^2}{7} = 31.43 = \text{trimmed sum of squares of deviations.}$$

From Table 2 we find that 31.43 should be divided by 18.935 to obtain an unbiased estimate of the variance of $\overset{\sqcup}{y}$. The standard error of $\overset{\sqcup}{y}$ is thus

$$\sqrt{\frac{31.43}{18.935}} = 1.29$$

I.

GENERAL CONSIDERATIONS

We shall use the following notations:

$$y_1 \leq y_2 \leq \ldots \leq y_n$$

are the ordered values in a sample;

$$\text{ave } \{ \ \}$$

indicates the average value, or expectation, of the expression in $\{ \ \}$;

$$\text{var } \{ \ \}$$

indicates the variance of the quantity following, as defined by

$$\text{var } u = \text{ave } (u^2) - (\text{ave } u)^2 \ ;$$

when clarity or precision require indication of the distribution from which the samples are drawn,

$$\text{ave}_N\{ \ \} \qquad \text{and} \qquad \text{var}_N\{ \ \}$$

will refer to averages and variances based on the standard normal distribution (average zero and variance units), while

$$\text{ave}_R\{ \ \} \qquad \text{and} \qquad \text{var}_R\{ \ \}$$

will refer to averages and variances based on the standard rectangular distribution (on the interval $0 \leq p \leq 1$).

The quantities of most interest to us will be denoted as follows, suppressing dependences on n, g, and the particular sample:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \text{untrimmed mean}$$

$$g = \text{number trimmed from each end,}$$

$$h = n - 2g = \text{number remaining,}$$

$$\overset{\cup}{y} = \frac{1}{n-2g}(y_{g+1} + y_{g+2} + \ldots + y_{n-g}) = \text{trimmed mean}$$

$$T = (y_{g+1} - \overset{\cup}{y})^2 + (y_{g+2} - \overset{\cup}{y})^2 + \ldots + (y_{n-g} - \overset{\cup}{y})^2$$

$$= {y_{g+1}}^2 + {y_{g+2}}^2 + \ldots + {y_{n-g}}^2 - (n - 2g)(\overset{\cup}{y})^2$$

$$= \text{trimmed sum of squares of deviations} = \text{TSSD.}$$

When we do need to bring in dependence on n and g, we shall often do this by writing $g + h + g$ as an argument. In such cases it will be understood that $g + h + g$ is the original sample size and that h is the trimmed sample size.

We shall also systematically let Σ^* refer to summation for i (or j) from $g + 1$ to $n - g$ (a total of $n - 2g = h$ values of i) over the same range. Then

$$\overset{\cup}{y} = \frac{1}{h}\Sigma^* y_i$$

$$T = \frac{1}{2h}\Sigma^*\Sigma^*(y_i - y_j)^2 = \Sigma^* {y_i}^2 - \frac{1}{h}\Sigma^*\Sigma^* y_i y_j$$

$$= \Sigma^* {y_i}^2 - \frac{1}{h}(\Sigma^* y_i)^2$$

as may easily be verified.

6. Relation to order-statistic moments.

The quantities that concern us most can be expressed in terms of order-statistic moments as follows:

$$\text{var } \overset{\cup}{y} = \text{ave}(\overset{\cup}{y})^2 - (\text{ave } \overset{\cup}{y})^2 = \text{ave}(\overset{\cup}{y})^2$$

$$= \frac{1}{h^2} \Sigma^* \Sigma^* \text{ ave}(y_i y_j) \text{ ,}$$

$$\text{ave } T = \Sigma^* \text{ ave}(y_i{}^2) - \frac{1}{h} \Sigma^* \Sigma^* \text{ ave } (y_i y_j),$$

$$\text{Div } (g + h + g) = \frac{\text{ave } T}{\text{var } \overset{\cup}{y}} = h^2 \; \frac{\Sigma^* \text{ ave } (y_i{}^2)}{\Sigma^* \text{ ave } (y_i y_j)} - h \text{ .}$$

Again we write $\text{Div}_N \, (g + h + g)$ or $\text{Div}_R(g + h + g)$ when needed for clarity or precision.

7. Normal distributions.

For the special case of sampling from a standard normal population, we can refer to Teichroew [3] for the values of $\text{ave}_N(y_i)$, $\text{ave}_N(y_i{}^2)$ and $\text{ave}_N(y_i y_j)$ for $n \leq 20$. (The corresponding variances and covariances are given by Sarhan and Greenberg a few pages later [2].)

Thus normal-theory variances of $\overset{\cup}{y}$'s and normal theory averages of TSSD's are easily available for normal samples of size no more than twenty. For example, the case of $17 = 6 + 5 + 6$, where a sample of 17 is trimmed to the central five observations, yields

$$\Sigma^* \Sigma^* \text{ ave } (y_{ij}) = 1.92257699$$

$$\Sigma^* \text{ ave } (y_i{}^*) = .674220047$$

whence

$$\text{var}_N \, \overset{\cup}{y} = \frac{1.92257699}{25} = .076903080$$

$$\text{Div } (6 + 5 + 6) = \frac{.674220047}{.076903080} - 5 = 3.7671397$$

Notice that if we had had an initial sample of 5, and had not trimmed it, the correct divisor would have been

$$\mathrm{Div}\,(0 + 5 + 0) = 20 \; .$$

Thus we must treat the sum of squares of deviations from a trimmed sample quite differently from the sum of squares of deviations from an untrimmed sample. This is emphasized by Table 3, which gives values of the ratio

$$\frac{\mathrm{Div}_N(0 + h + 0)}{\mathrm{Div}_N\,(g + h + g}$$

of the divisors which are appropriate on normal theory in the two cases.

TABLE 3

Values of $\dfrac{\mathrm{Div}_N(0 + h + 0)}{\mathrm{Div}_N(g + h + g)}$

Size of sample after trimming	$\frac{n-h}{2}$ = g = number of observations trimmed from each end									
h	0	1	2	3	4	5	6	7	8	9
1	-	-	-	-	-	-	-	-	-	-
2	1	1.982	2.979	3.979	4.980	5.982	6.983	7.984	8.985	9.986
3	1	1.895	2.794	3.694	4.594	5.494	6.394	7.295	8.195	
4	1	1.810	2.611	3.411	4.210	5.009	5.809	6.608	7.407	
5	1	1.736	2.455	3.170	3.883	4.596	5.309	6.022		
6	1	1.675	2.326	2.970	3.612	4.253	4.894	5.535		
7	1	1.624	2.218	2.804	3.387	3.969	4.550			
8	1	1.580	2.127	2.665	3.198	3.730	4.261			
9	1	1.543	2.050	2.546	3.038	3.528				
10	1	1.511	1.984	2.445	2.901	3.355				
11	1	1.483	1.926	2.357	2.782					
12	1	1.458	1.876	2.280	2.679					
13	1	1.436	1.831	2.212						
14	1	1.417	1.791	2.152						
15	1	1.399	1.756							
16	1	1.383	1.723							
17	1	1.368								
18	1	1.355								
19	1									
20	1									

Thus changes in this factor with shape deserve exploration. For this purpose, exploration of shapes far more extreme than are likely to arise in practice is reasonable since the aim is to discover magnitude of dependence rather than to balance losses.

For this purpose, the accessibility of order statistic moments for rectangular distributions is convenient and useful, since the rather extreme shape of the rectangular distribution is not a handicap.

It is shown in Section 11 that, for a rectangular distribution of unit length (which will serve us as well as any other as a standard rectangular distribution) that if g values are deleted from each tail of a sample size $n = h + 2g = g + h + g$, leaving h central values for the computation of the trimmed mean and the trimmed sum of squares of deviations, then:

rectangular variance of trimmed mean =

$$\text{var}_R(\overset{\sqcup}{y}) = \frac{1}{4(n+2)} \left(1 - 2\,\frac{h^2 - 1}{3h(n+1)}\right)$$

average of trimmed sum of squares of deviations =

$$\text{ave}_R T = \frac{(h+2)\,(h+1)\,(h-1)}{12\,(n+2)\,(n+1)}$$

reciprocal of divisor for conversion

$$\frac{1}{\text{Div}_R(g+h+g)} = \frac{3(n+1)}{(h+2)\,(h+1)\,(h-1)} - \frac{2}{(h+2)h} \quad .$$

Multiplication of the values already obtained for the normal-theory conversion-divisor conversion by the reciprocal of the rectangular-theory conversion-divisor yields the values of ratios of divisors set forth in Table 4. It is clear from this table that the normal theory conversion-divisor is in any case approximately equal to the rectangular-theory conversion-divisor, and, as would be expected, the approximation becomes better as the amount of trimming increases.

Table 4

Values of $\frac{\mathrm{Div}_N(g + h + g)}{\mathrm{Div}_R(g + h + g)}$

h = size of trimmed samples	g = number of observations trimmed from each end									
h	0	1	2	3	4	5	6	7	8	9
2	1	1.0092	1.0071	1.0053	1.0040	1.0031	1.0024	1.0020	1.0017	1.0014
3	1	1.0025	1.0021	1.0017	1.0013	1.0011	1.0009	1.0007	1.0006	
4	1	.99470	.99570	.99678	.99756	.99811	.99850	.99878	.99899	
5	1	.98733	.98908	.99142	.99325	.99460	.99560	.99637		
6	1	.98079	.98271	.98600	.98873	.99082	.99241	.99363		
7	1	.97513	.97682	.98074	.98419	.98693	.98907			
8	1	.97028	.97146	.97575	.97977	.98304	.98566			
9	1	.96615	.96662	.97109	.97551	.97923				
10	1	.96263	.96227	.96676	.97146	.97554				
11	1	.95963	.95838	.96275	.96763					
12	1	.95708	.95491	.95907	.96403					
13	1	.95490	.95180	.95568						
14	1	.95304	.94903	.95256						
15	1	.95145	.94653							
16	1	.95010	.94431							
17	1	.94894								
18	1	.94795								
19	1									
20	1									

9. Future work.

Besides the questions of (i) relative efficiency for reasonable population shapes, (ii) allowance for selection bias when the amount of trimming is allowed to vary from sample to sample, and (iii) improvement from an unbiased-estimate-of-variance procedure to a confidence procedure, all of which are very important to the practical use of "trimmed" techniques, the considerations of later sections about the rectangular case make it clear that normal theory and rectangular theory can be usefully compared for other sorts of "trimmed" procedures. The mid-range (mean of highest and lowest values) of the trimmed sample needs to be considered as an indicator of location. It is, of course, an inner (or quasi-) midrange of the entire sample. For both trimmed means and inner midranges it is appropriate to consider at least the following as bases for estimating variability:

(a) sum of squares of deviations for the same trimmed sample,

(b) square of the range of the same trimmed sample,

(c) sum of squares of deviations for a less vigorously trimmed sample

(d) square of the range of a less vigorously trimmed sample.

It is hoped to report on some of these shortly.

II.

DERIVATIONS, DISCUSSIONS, DETAILS

10. Empirical behavior of Div (g + h + g).

When the normal-theory behavior of Div (g + h + g) was examined, it was noticed that, for h fixed and g changing the first differences of its reciprocal decreased somewhat for $h > 3$, increased slightly for $h = 2, 3$, but in both cases rapidly became constant as g increased. This is illustrated, for two values of h, in Table 5. This observation immediately makes it possible to extend the tables of divisors beyond total sample size 20 by empirical extrapolation. Such a process could be used to quite good effect without further support. However, its use will be simpler, and somewhat more precise, if it can borrow support from algebraic considerations which apply either to some other distribution shape or in some asymptotic sense. Results for the rectangular case are easily obtained, and may be shown to hold asymptotically for all distributions smooth at the median.

Table 5

ILLUSTRATION OF APPROACH OF G-WISE DIFFERENCES OF RECIPROCALS OF DIVISORS

		h = 2			h = 5	
g	$\frac{1}{\text{Div}_N(g + 2 + g)}$	δ_g	$\delta(\delta_g)$	$\frac{1}{\text{Div}_N(g + 5 + g)}$	δ_g	$\delta(\delta_g)$
0	.50000			.05000		
		.49088			.072277	
1	.99088		+.00763	.12277		-.03704
		.49851			.03573	
2	1.48939		+.00165	.15850		-.00006
		.50016			.03567	
3	1.98955		+.00043	.19417		-.00003
		.50059			.03564	
4	2.49014		+.00008	.22981		-
		.50067			.03564	
5	2.99081		-.00003	.26545		-
		.50064			.03564	
6	3.49145		-.00006	.30109		
		.50058				
7	3.99203		-.00005			
		.50053				
8	4.49256		-.00006			
		.50047				
9	4.99303					

11. The rectangular case.

The distributions of order statistics of samples from the standard rectangular distribution are well-known [1] as are expressions for their moments. If p_i and p_j are the ith and jth order statistics of a sample of n from the standard rectangular, where $i \leq j$, then

$$\text{ave } (p_j - p_i)^2 = (\text{ave } y_j - \text{ave } y_i)^2 + \text{var } y_j - 2\text{cov } (y_j,\ y_i) + \text{var } y_i$$

$$= \left(\frac{j}{n+1} - \frac{i}{n+1}\right)^2 + \frac{1}{(n+1)^2(n+2)} \left(j\ (n-j+1) - 2i(n-j+1) + i\ (n-i+1)\right)$$

This is a function of n and $j-i$ alone, and hence equal to ave $(p_{j-i})^2$, as would be expected from the symmetric distribution of equivalent blocks [4]. What is important next is that $(n+1)(n+2)\text{ave}(p_j-p_i)^2$ depends only on $j-i$. As i and j run over any h consecutive indices of a sample of n, the values of $j-i$ are exactly the same, and occur with the same multiplicity, as if i and j ran through a sample of size h. Consequently $(n+1)(n+2) \sum_i^* \sum_j^* \text{ave } (p_j - p_i)^2 = (h+1)(h+2) \sum \sum \text{ave } (p_j - p_i)^2$ where Σ^* is over some h consecutive values of a sample of n and Σ is over <u>all</u> values from 1 to h of a sample of h.

Let now $p_1 \leq p_2 \leq p_3 \leq \cdots \leq p_h$ be the order statistics of a sample of h (not n) from the standard rectangular, and let $p_1^* \leq p_2^* \leq p_3^* \leq \cdots \leq p_n^*$ be the order statistics of an independent sample of n (not h) from the same distribution. Let $T(p)$ and $T(p^*)$ be the corresponding TSSD's, in the first case for all h values and in the second case for the central h values. Since

$$2h \cdot T(p^*) = \Sigma^*\Sigma^*(p_j - p_i)^2$$

$$2h \cdot T(p) = \Sigma\ \Sigma\ (p_j - p_i)^2$$

we must have

$$\begin{aligned}\text{ave}_R\, T(p^*) &= \frac{1}{2h}\,\text{ave}_R\, \Sigma^*\Sigma^*\,(p_j{}^* - p_i{}^*)^2 \\ &= \frac{1}{2h(n+1)(n+2)}\,\Sigma^*\Sigma^*\,\text{ave}_R(n+1)(n+2)(p_j{}^* - p_i{}^*)^2 \\ &= \frac{1}{2h(n+1)(n+2)}\,\Sigma\,\Sigma\,\text{ave}_R(h+1)(h+2)(p_j - p_i)^2 \\ &= \frac{(h+1)(h+2)}{(n+1)(n+2)}\,\text{ave}_R\, T(p) = \frac{(h-1)(h+1)(h+2)}{12(n+1)(n+2)}\end{aligned}$$

for, since $T(p)$ is an untrimmed sum of squares for a sample of size h, ave $T(p) = (h-1)\sigma^2$ for any distribution.

Now let us turn to var $\overset{\sqcup}{p}$. Recall that, for $i < j$

$$\begin{aligned}(n+1)^2(n+2)\,\text{cov}(p_i,p_j) = i(n+1-j) &= \left(\frac{n+1}{2} + c\right)\left(\frac{n+1}{2} - d\right) \\ &= \left(\frac{n+1}{2}\right)^2 - \left(\frac{n+1}{2}\right)(d - c) - cd\end{aligned}$$

where $2i = (n+1) + 2c$, $2j = (n+1) + 2d$, so that c and d range over h values with average zero and unit spacing between adjacent values. Hence

$$(n+1)^2(n+2)\,\underset{i}{\Sigma^*}\underset{j}{\Sigma^*}\text{cov}(p_j,p_i) = h^2\left(\frac{n+1}{2}\right)^2 - 2\,\frac{n+1}{2}\,\underset{|d-c|}{\Sigma^*}\,(h - |d-c|)|d-c|$$

since $\Sigma^*\Sigma^*\, cd = (\Sigma^* c)\,(\Sigma^* d) = 0\cdot 0 = 0$.

Now

$$\begin{aligned}\sum_{k=1}^{h}(h-k)\cdot k = h\,\Sigma\, k - \Sigma k^2 &= \frac{h\cdot h(h+1)}{2} - \frac{h(h+1)(2h+1)}{6} \\ &= \frac{h(h+1)}{6}(3h - 2h - 1) = \frac{(h+1)(h)(h-1)}{6} = \binom{h+1}{3}\end{aligned}$$

so that

$$\begin{aligned}\text{var}\,\overset{\sqcup}{p} = \frac{1}{h^2}\Sigma^*\Sigma^*\,\text{cov}(p_j,p_i) &= \frac{1}{h^2(n+1)^2(n+2)}\left[h^2\,\frac{(n+1)^2}{4} - 2\,\frac{n+1}{2}\binom{h+1}{3}\right] \\ &= \frac{1}{4(n+2)}\left[1 - 2\,\frac{h^2-1}{3h(n+1)}\right]\end{aligned}$$

(for $h = 1$, this checks with the variance of the median $\overset{\prime}{p}$, namely $1/4(n+2)$, while for $h = n$ it reduces to $1/12n$, as it should.

The conversion divisor is thus

$$\mathrm{Div}_R\,(g + h + g) = \frac{(h+2)(h+1)(h-1)}{12(n+2)(n+1)} \Big/ \frac{3h(n+1) - 2(h^2 -1)}{12h(n+1)(n+2)}$$

$$= \frac{(h+2)(h+1)(h)(h-1)}{3h(n+1) - 2(h^2-1)} = \frac{(h+2)(h+1)(h)(h-1)}{3h\cdot n - (2h^2 -3h-1)}$$

which reduces to $(n)(n-1)$ when $h=n$, as it should. Its reciprocal can be written

$$\frac{1}{\mathrm{Div}_R(g + h + g)} = \frac{3(n+1)}{(h+2)(h+1)(h-1)} - \frac{2}{(h+2)(h)}$$

which is obviously linear in n.

If we fix h, and let g increase unit by unit, n will increase in steps of 2, and the rectangular theory reciprocal will increase in steps of

$$\frac{6}{(h+2)(h+1)(h-1)} .$$

12. <u>The asymptotic case.</u>

Consider next the case of an arbitrary distribution where h is fixed and n is large. If $y = r(p)$ is the representing function of the distribution, so that $F(r(p)) = p$ where $F(y)$ is the corresponding cumulative, then

$$y_i = r(p_i)$$

where $y_1, y_2, \ldots, y_n$ are the order statistics of a sample of n y's and $p_1, p_2, \ldots, p_n$ are the order statistics of the corresponding sample of n p's.

Put $q' = p_g$ and $q'' = p_{n+1-g}$, so that the h central p_i fall between q' and q''. It is a consequence of Wald's principle [4] that, conditional upon the values of q' and q'', these h p_i's are distributed like a sample of h from the rectangular distribution with q' and q'' as end point. Conditional on q' and q'' we have the following averages and variances:

$$\text{ave } (\breve{p} \mid q', q'') = \frac{1}{2} (q' + q'')$$

$$\text{var } (\breve{p} \mid q', q'') = \frac{(q'' - q')^2}{12h}$$

$$\text{ave } (T(p) \mid q', q'') = \frac{h-1}{12} (q'' - q')^2$$

whence

$$\begin{aligned} \text{var}_R\, \breve{p} &= \underset{q',\, q''}{\text{ave}}\ \text{var } (\breve{p} \mid q', q'') + \underset{q',\, q''}{\text{var}}\ \text{ave } (\breve{p} \mid q', q'') \\ &= \frac{1}{12n}\ \text{ave } (q'' - q')^2 = + \frac{1}{2}\ \text{var } (q'' + q') \end{aligned}$$

and

$$\text{ave}_R\, T(p) = \frac{h-1}{12}\ \text{ave } (q'' - q')^2$$

so that the reciprocal of the conversion factor satisfies

$$\frac{1}{\text{Div}_R(g + h + g)} = h(h-1)\, \frac{\text{var } \breve{p}}{\text{ave } T(p)} = 1 + 6h\ \frac{\text{var } (q'' + q')}{\text{ave } (q'' - q')^2} \quad .$$

If now $z' = y_g$ and $z'' = y_{n+1-g}$, so that the h central y_i fall between z' and z'', it again follows from Wald's principle that, conditional onthe values of z' and z'', these y_i are distributed like a sample from whatever may be the distribution of y truncated onto the interval from z' to z''.

If n is very large, the interval from z' to z'' will be short and will lie close to the median of the distribution of y . If that distribution is smooth near its median, the result of truncating it onto any small interval near the median will be very nearly a rectangular distribution.

Hence

$$\frac{1}{\mathrm{Div}_D(g+h+g)} \quad h(h-1) \quad \frac{\mathrm{var}\ \overset{\sqcup}{y}}{T(y)} \approx 1 + 6h \frac{\mathrm{var}\ (y'' + y')}{\mathrm{ave}\ (y'' - y')^2}$$

where D stands for any distribution smooth near the median, and T(y) is the TSSD for the y's. Moreover,

$$z' = r(q') \quad \text{and} \quad z'' = r(q'')$$

where r will behave very nearly linearly, so that

$$\frac{\mathrm{var}\ (y'' + y')}{\mathrm{ave}(y'' - y')^2} \approx \frac{\mathrm{var}\ (q'' + q')}{\mathrm{ave}\ (q'' - q')^2}$$

consequently

$$h(h-1) \ \frac{\mathrm{var}\ \overset{\sqcup}{y}}{\mathrm{ave}\ T(y)} \approx 1 + 6h \frac{\mathrm{var}(q'' + q')}{\mathrm{ave}(q'' - q')^2} = h(h-1) \ \frac{\mathrm{var}\ \overset{\sqcup}{p}}{\mathrm{ave}\ T(p)}$$

and we see that asymptotically, for fixed h and very large n , the value of the conversion factor will not depend upon the shape of the parent distribution, so long as that distribution is smooth near the median.

If the distribution of y is symmetric, then

$$r(p) = a + b(p - \tfrac{1}{2}) + d(p - \tfrac{1}{2})^3 + \ldots$$

and deviations from linearity are of order $(p - \frac{1}{2})^2$ times the linear deviations. Since $(p - \frac{1}{2})^2$ is of order 1/n , the fractional deviations of the conversion factor for any two symmetrical distributions from one another are at most of order $1/n$ for each fixed h .

Suppose that, for two symmetrical parent distributions, the conversion factors for some h satisfy:

$$\text{factor}^{-1} = A_1 \cdot (n + 1) + B_1 + C_1(n)$$
$$\text{factor}^{-1} = A_2 \cdot (n + 1) + B_2 + C_2(n)$$

where $C_1(n)$ and $C_2(n)$ both tend to zero as n increases. Their ratio can only approach unity as n tends to infinity if $A_1 = A_2$. For the standard rectangular distribution

$$\text{factor}^{-1} = \frac{3}{(h+2)(h+1)(h-1)} (n+1) - \frac{2}{(h+2)h} \quad .$$

Consequently, for any symmetrical distribution for which the general form applies,

$$\text{factor}^{-1} = \frac{3}{(h+2)(h+1)(h-1)} (n+1) + \text{constant} + C(n)$$

where $C(n)$ tends to zero, while the difference between the reciprocals of the factor for n and $n-1$ will be

$$\delta(\text{factor}^{-1}) = \frac{3}{(h+2)(h+1)(h-1)} + [C(n) - c(n-2)].$$

13. Suggested alternatives.

The discussion of the last paragraph shows, upon reexamination, that the reason why the conversion factor does not depend upon h alone lies in the ratio

$$\frac{\text{var [average of distribution trimmed to } (z', z'')]}{\text{ave [variance of distribution trimmed to } (z', z'')]} \quad .$$

Thus it appears that perhaps the most natural way to build in some compensation is to use as a basis for estimating the trimmed variance of the mean of h values, not the TSSD or squared range of the same h central values, but the TSSD or squared range of a greater number of values, perhaps 1 + h + 1 or 2 + h + 2 or 3 + h + 3 . These possibilities will be considered numerically in a later report.

REFERENCES

[1] M. G. Kendall and Alan Stuart 1958; The advanced theory of statistics, Vol. 1; London, Charles Griffin (esp. §11.4ff).

[2] A. E. Sarhan and B. G. Greenberg 1956; Estimation of location and scale parameters by order statistics from singly and doubly censored populations. I; 27 Annals Math. Statist. 427-451.

[3] D. Teichroew 1956; Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution; 27 Annals Math. Statist. 41--426.

[4] John W. Tukey 1947; Non-parametric estimation II. Statistically equivalent blocks and tolerance regions - the continuous case; 18 Annals Math. Statist. 529-539.

[5] John W. Tukey 1960; A survey of sampling from contaminated distributions; Chap. 39 (pages 448-485) in Contributions to Probability and Statistics; (Ed. Olkin et al), Stanford, University Press.